Use of Artificial Intelligence in Diagnosis of Head and Neck Precancerous and Cancerous Lesions: A Systematic Review

ABSTRACT

This systematic review analyses and describes the application and diagnostic accuracy of Artificial Intelligence (AI) methods used for detection and grading of potentially malignant (pre-cancerous) and cancerous head and neck lesions using whole slide images (WSI) of human tissue slides. Electronic databases MEDLINE via OVID, Scopus and Web of Science were searched between October 2009 - April 2020. Tailored search-strings were developed using database-specific terms. Studies were selected using a strict inclusion criterion following PRISMA Guidelines. Risk of bias assessment was conducted using a tailored QUADAS-2 tool. Out of 315 records, 11 fulfilled the inclusion criteria. AI-based methods were employed for analysis of specific histological features for oral epithelial dysplasia (n=1), oral submucous fibrosis (n=5), oral squamous cell carcinoma (n=4) and oropharyngeal squamous cell carcinoma (n=1). A combination of heuristics, supervised and unsupervised learning methods were employed, including more than 10 different classification and segmentation techniques. Most studies used uni-centric datasets (range 40-270 images) comprising small sub-images within WSI with accuracy between 79-100%. This review provides early evidence to support the potential application of supervised machine learning methods as a diagnostic aid for some oral potentially malignant and malignant lesions; however, there is a paucity of evidence using AI for diagnosis of other head and neck pathologies. Overall, the quality of evidence is low, with most studies showing a high risk of bias which is likely to have overestimated accuracy rates. This review highlights the need for development of state-of-the-art deep learning techniques in future head and neck research.

Keywords

Artificial intelligence; machine learning; head and neck cancer; oral cancer; pre-cancer; oral potentially malignant disorders, dysplasia, squamous cell carcinoma, deep learning, systematic review.

INTRODUCTION

Head and neck cancers (HNC) encompass a large group of cancers, most commonly squamous cell carcinomas (SCC) (90%) of the oral cavity, nasal cavity, sinuses, salivary glands, pharynx and larynx. Primary risk factors include tobacco and betel nut use[1], alcohol consumption[2], radiation[3], immunodeficiency[4] and specific viruses including Human Papillomavirus (HPV) 16 and 18 (for oropharyngeal squamous cell carcinoma, OPSCC)[5] and Epstein-Barr virus (for nasopharyngeal squamous cell carcinoma, NPSCC)[6,7]. Chronic exposure to these carcinogenic factors and/or infection status can result in dysplastic changes in the oral, oropharyngeal, nasal or nasopharyngeal mucosa, which may lead to the development of HNC. The incidence of HNC continues to rise, making it the sixth leading group of cancers worldwide[8,9]. In 2018, HNC accounted for more than 650,000 new cases and 33,000 deaths annually worldwide[10]. In the UK, the number of new patientshas increased by 22% over the last decade, with almost 12,000 new diagnoses every year (33 every day)[11].

Despite advancements in medical and surgical techniques, prognosis of HNC remains poor with a five-year survival rate between 28-67%[11]. Due to late presentation, even successful treatment of HNC is associated with multiple functional problems including masticatory, speech and swallowing impairments which can significantly reduce the quality of life[12]. Early diagnosis of potentially malignant head and neck lesions can prevent cancer development in up to 88% of cases[11], however most patients are diagnosed at a late stage of disease (62% diagnosed at stage III or IV)[13]. The conventional diagnosis of suspicious head and neck lesions

involves clinical, radiological and histopathological assessment. The latter is the gold standard providing important prognostic information (i.e. grade for dysplasia and cancers) which can guide clinical treatment decisions[14,15]. However, histological interpretation can be subjective with differences in interpretation[16], variation in consistency[17] and may not provide effective risk stratification or management guidance. This highlights the importance of novel methods and technologies for more consistent, efficient and accurate diagnosis to aid clinical decision-making and to improve HNC related patient survival.

Over the past decade, Artificial Intelligence (AI) has gained popularity in cancer research where it has been shown to increase diagnostic accuracy and efficiency by providing quantifiable outputs to predict cancer behaviour and prognosis[18,19,20]. Machine learning (ML), a branch of AI, has been shown to reduce variability in grading of dysplasia and cancers by ensuring standardisation and consistency in addition to informing treatment decisions[21]. ML uses computational methods to 'learn' information and patterns directly from data. This learning can be *supervised (*involving training of ML models on a known data input and output i.e. histology slides with associated diagnostic annotations) or *unsupervised (*which involves mining and extraction of hidden patterns from input data without any pre-defined information). ML algorithms adaptively improve their performance with an increasing number of 'learning or training' samples, enabling the computer to essentially 'learn from experience'. Classical supervised ML approaches include *semantic segmentation* and *classification. Segmentation* involves dividing high-resolution digital whole slide images (WSI) of human tissue into regions of clinical relevance followed by deconstruction of the WSI into smaller patches (sub-images) by a process known as 'patch extraction'. This enables ML algorithms to compute local and global features which can be explored for significance during the 'learning or training' phase. *Classification* involves organising and classifying new observations based on specific attributes (e.g. morphology of nuclei) learnt from previous data input. Both techniques are commonly used approaches in cancer research, providing useful diagnostic and prognostic outputs. Other relevant computational pathology terms have been described in Table 1.

With the evolution of computational power and image analysis algorithms, there is now an increasing amount of evidence demonstrating the success of AI-based image analysis from WSI of human tissue slides[22]. Several studies have demonstrated the potential for AI-based methods to reliably predict diagnosis, prognosis, mutational status, and response to treatment in a range of cancers including colorectal, lung, skin and breast malignancies[23,24,25,26,27]. These studies highlight the potential for AI-based methods in provision of faster, consistent, accurate and reproducible information regarding cancer diagnosis and prognosis which can complement the conventional (and largely subjective) light microscopy analysis by experienced pathologists.

The alarming rise in global HNC incidence and its poor prognosis makes it ideally suited for application of AI-based methods to aid objective diagnosis and provide valuable prognostic information. We performed a systematic review of literature published in the last ten years, to assess the application and diagnostic accuracy of AI/ML methods for detection and grading of potentially malignant and cancerous head and neck lesions. To the best of the authors' knowledge, this is the first study reviewing the use and application of AI-based methods for head and neck lesions.

MATERIALS AND METHODS

The systematic review was conducted using a predetermined protocol which followed the recommendations of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and checklist[28]. The protocol was registered in the International Prospective Register of Systematic Review (PROSPERO) database, CRD42019153023.

Outcome definitions

The primary outcomes were the specific histopathological features used for diagnosis and grading of the head and neck lesion under study, in addition to the methods and performance of the proposed AI/ML techniques. Descriptive analysis was conducted on these outcomes and the reported diagnostic performance measures (i.e. sensitivity, specificity, accuracy, F1-score) where possible.

Literature search

Electronic databases search of MEDLINE via OVID, Scopus and Web of Science was conducted to retrieve articles published in the English language between October 2009 and April 2020. The Cochrane library was also consulted. This time period was chosen due to the rapid evolution of AI methods and their application to cancer diagnostics over the last decade.

The search strategy was jointly developed by the multidisciplinary authorship team in collaboration with a medical information specialist (librarian from University of Sheffield, UK). Input from an expert Oral and Maxillofacial Pathologist (SAK, University of Sheffield, UK) and a Professor of Computational Pathology (NMR, University of Warwick, UK) ensured adequate selection of clinical and technical terms and controlled vocabulary for optimal identification of articles.

Tailored search strings containing keywords and database-specific medical subject headings (MeSH) for the two major topics (AI/ML and potentially malignant or cancerous head and neck disorders/lesions) were developed. Multiple variations of search terms were combined to produce different sets of results and the final search strategy was pilot-tested and modified accordingly (Appendix 1). Grey literature and the reference lists of selected articles were also screened for relevant studies that may not have been identified through the database searches. The electronic databases search was conducted with assistance from an experienced Librarian at the University of Sheffield, UK. Article citations were exported to EndNote® reference manager software (Clarivate Analytics, Philadelphia, USA) and duplicate articles were removed.

Study selection

Two independent reviewers (HM, MS) retrieved the literature, and screened titles and abstracts. Where insufficient information was available to determine eligibility, the full report was obtained for further assessment. Articles that did not meet the eligibility criteria were excluded. In the second stage of study selection, the same two reviewers (HM, MS) independently assessed the full-text reports to obtain a shortlist of relevant articles. The shortlists were compared, and differences discussed, obtaining a final selection of studies. In case of any discrepancies in article selection, a discussion with senior members of the review team (NMR, SAK) took place to reach a mutual final decision. For relevant articles with overlapping datasets or results, the most recent publication was included.

The following criteria were applied for the selection of eligible studies for this review:

Inclusion criteria:

- Studies using AI for automated detection, grading and classification of potentially malignant and cancerous head and neck disorders/lesions.
- Studies exploring diagnostic accuracy of the applied AI/ML method providing sensitivity, specificity, accuracy, F1-scores as outcome measures.
- Studies published in indexed journals between October 2009 - April 2020.

Exclusion criteria:

- Studies not using WSI of human tissue slides.
- Studies not using histological image modalities (e.g. radiographic, photographic, cytology, genomic data etc.).
- Studies using AI/ML to predict disease progression, prognosis, metastasis, recurrence, survival or treatment efficacy (i.e. those not primarily investigating detection, grading and classification of head and neck lesions).
- Studies using AI/ML for detection and diagnosis of thyroid or oesophageal cancer.
- Narrative reviews, letters to editors, commentaries and conference abstracts.
- Studies not available in the English language.

Data extraction

Relevant data from selected articles was extracted, processed and tabulated into a pre-developed data collection form in Microsoft Excel® (Microsoft Corporation, Washington, USA) by two reviewers (HM, MS). The following information was recorded:

- Study details (authors, year and country of publication, aims)
- Study methods (design, dataset size and selection)
- Description of outcome variables (AI/ML methods used, head and neck lesion and histological parameter under study, training and validation sample details) and its outcome measures (reported diagnostic accuracy, effect measures)
- Other relevant details (funding information, sources of support, conflict of interest disclosure)

Methodological quality and risk of bias assessment

The methodological quality of individual studies and risk of bias was assessed using the Quality Assessment of Studies of Diagnostic Accuracy – Revised QUADAS-2 tool[29]. This tool is designed specifically for use in systematic reviews to evaluate the risk of bias and applicability of the primary diagnostic accuracy studies. The tool was adapted with input from an Oral and Maxillofacial Pathologist (SAK) and a Professor of Computational Pathology (NMR) to ensure relevant signalling questions were included to reliably and fairly assess the quality of included studies in relation to: 1) sample selection 2) index test and 3) reference standard. The tailored QUADAS-2 tool was piloted on five studies by two independent reviewers (HM, MS) and differences were resolved with consensus. The overall score for each study was determined by combining the number of satisfied criteria, with a higher score representing higher methodological quality. The outcome of the methodological quality assessment is presented graphically in Table 3 and the influence of bias risk on our results was discussed where applicable.

Data synthesis

A narrative synthesis of the main study findings is presented. Due to the large variation in outcome definitions and heterogeneity of retrieved data, a meta-analysis for calculation of adjusted pool estimates was not carried out.

RESULTS

Search results

The electronic database search retrieved a total of 314 articles (MEDLINE via OVID: 154, Scopus: 81 and Web of Science: 79). In addition, one article was identified through citation searching. Following removal of duplicate studies, 288 articles were selected. After the first screen based on title and abstract, 259 articles did not satisfy the inclusion criteria and were therefore excluded. A comprehensive full-text examination of

the remaining 29 articles excluded a further 18; resulting in 11 eligible articles for inclusion in this review paper (Figure 1).

Out of 315 articles, 277 were excluded, with a large proportion not eligible due to the use of imaging modalities other than histology (n=157). Many studies did not address the research question directly (n=67) as they focussed on the application of AI algorithms to predict disease prognosis, recurrence, metastasis or treatment success. Other reasons for exclusion included studies which did not use AI based methods (n=52) or human tissue (n=1).

Description of studies

Table 2 summarises the main findings for the included studies. In six studies, AI-based methods were used to detect oral potentially malignant disorders (OPMD)[52,54,55,57,58,62] with five of these focussing on the detection of oral submucous fibrosis (OSF) specifically. Four studies aimed to detect oral squamous cell carcinoma (OSCC)[53,56,59,60] and one study aimed to classify OPSCC[61]. Overall, seven studies were conducted in India[53,54,55,57,58,59,62], two in China[60,61], one in USA[52] and one in Germany[56]. Eight studies were published between 2009 and 2015[52,54,55,57,58,60,61,62] and three were published after 2015[53,56,59].

Results of the selected studies have been presented based on the type of head and neck lesion being analysed, which includes OPMD, OSCC, and OPSCC. Following this, the methods used in the selected studies will be presented, which will describe the type of AI/ML technique, dataset sample and the diagnostic performance for each.

Detection of OPMD

Baik *et al.*[52] quantified nuclear phenotypic changes in oral epithelial dysplasia (OED) lesions using an automated nuclear phenotypic score (a-NPS). The a-NPS was used to classify suspicious oral lesions based on the risk of progression to OSCC. The tissue samples used for algorithm training compared relatively normal oral mucosa (i.e. amalgam tattoo or melanotic macule, 34%) to OSCC from high risk intra-oral sites (floor of mouth and lateroventral tongue, 66%). Following training, the algorithm was tested on biopsies diagnosed as hyperplasia, mild or moderate dysplasia including almost an equal representation of transformed and untransformed lesions. The study used a robust experimental design to produce good accuracy (78% sensitivity and 71% specificity) highlighting the a-NPS as a potentially useful prognostic adjunct.

Five other studies focussed on detection of OSF[54,55,57,58,62] using a variety of supervised ML methods to differentiate between normal tissue, OSF with dysplasia or atrophy and OSF without dysplasia or atrophy (Table 2). Krishnan *et al*.[57] classified the number of sub-epithelial connective tissue (SECT) cells (excluding endothelial cells) in oral mucosa of normal and OSF tissue. Specific histological features including SECT cell shape, size and dimensions were evaluated with focus on round shaped cells (e.g. macrophages, lymphocytes, mast cells and neutrophils) and spindle shaped cells (fibroblasts, fibrocytes, histiocytes and endothelial). In addition, geometric properties such as the compactness and eccentricity of these cells were considered for classification. The results demonstrated a classification accuracy of 88.69%, although this was based on a small dataset. In another study, Krishnan *et al*.[58] used a texture-based method for segmentation of the constituent layers of the epithelium in OSF tissue (based on density and thickness of individual layers) to distinguish it from normal tissue. The standard performance measures were not reported in this study (Table 2).

Detection of OSCC

Das *et al.*[53] aimed to detect OSCC using a two-stage approach. This involved segmentation of constituent layers of the oral mucosa (into epithelial, subepithelial and keratin layers) followed by texture-based classification of keratin pearls from segmented keratin regions. The detection accuracy for keratin pearls was

reported as 96.88% however this was based on a small dataset comprising small patches (sub-images) within WSI.

Rahman *et al.*[59] used a texture-based classifier to distinguish between normal and cancerous cells achieving an accuracy of 100% using small patches within WSI. In another study, Sun *et al.*[60] developed an automated colour-based feature extraction system to segment and classify OSCC stained with anti-CD34 antibody. Specific histological features (vessel area/number/density and nuclei area/number) were computed to enable quantitative differentiation between different OSCC stages. Results demonstrated sensitivities of 49.11%, 64.17%, 58.55%, 79.60% for OSCC stages I-IV, respectively.

Detection of OPSCC

Fouad *et al*.[61] used unsupervised ML methods for automated identification of specific tissue compartments (cells and nuclei) in OPSCC tissue microarrays. Measurements of cell and nuclei colour and morphology were used for classification of epithelial and stromal tissue. This study compared their results with other standard segmentation methods and reported relatively low recognition accuracy (pixel-level F1 score of 80-81%) attributed to the lack of pre-defined manual annotations often used in supervised learning methods.

ML methods used in selected studies

ML algorithms can be divided into two groups: *classical* and *modern*. The classical methods require small amounts of training data and computational resources for pattern recognition in comparison to modern methods. However, modern methods often outperform classical methods in addressing most ML problems. Deep learning is a modern ML approach, in which algorithms mimic the brain's neural networks to learn without supervision however it can suffer from the 'black box' problem, unlike classical ML methods which are easier to interpret. A hierarchical classification of ML methods used in the selected studies is presented in Figure 2.

In most of the selected studies, classical supervised ML approaches have been used although three classical unsupervised methods have also been employed, including Otsu and Watershed (for image segmentation into two or more classes) and Clustering (e.g. K-Mean and Agglomerative Hierarchical Clustering). The most frequently applied ML methods were from the classical supervised group, which included nine different techniques (Figure 3). The majority of these supervised methods belong to the handcrafted feature-based classical ML group, although in four studies[53,54,55,62] modern ML methods (neural networks) were employed. These nine methods differ significantly in their learning strategies, as outlined below:

- *Sugeno Fuzzy*[30] involves ML of fuzzy rules from the training dataset.
- *Decision Tree*[31] generates a binary tree based on training features for classification.
- *Random Forest*[32] builds a classification model using a set of decision tree-based classifiers.
- *K-Nearest Neighbour*[33] classifies an input image based on its similarity with other training set images, which enables the most dominant class of K to be assigned to the input image.
- *Bayesian Classifier*[34] use the Bayes rules to calculate the probability of an input sample to be a member of a specific class where the final label is assigned to the most probabilistic class for the given input image.
- *Linear Discriminant Analysis*[35] learns a linear combination of the features from training images to predict the label of test images.
- *Support Vector Model (SVM)*[36] learns a set of parameters from the training image to find a hyperplane which splits the training images into two classes. Same parameters are then used to classify test images.

- *Gaussian Mixtures Model*[37] learns multiple models from the training images to classify it into multiple classes.
- *Neural Network*[38] methods learn the representation of the training images using a gradient descent-based learning method. These methods require large training datasets compared to other aforementioned supervised based learning methods.

A combination of classification and segmentation methods were used to detect potentially malignant and cancerous head and neck lesions (Figure 3). Overall, five studies used AI based classifiers[52,54,56,59,62], four studies used segmentation methods[53,58,60,61] and two studies used a combination of classification and segmentation methods[55,57]. The most frequently used methods included SVM[36], Neural Network[38], Random Forest[32] and clustering. In the majority of studies, multiple methods were used for intermediate and final stages of the proposed AI framework, although there were considerable variations in the overall number of methods used between studies. For example, in one study Krishnan *et al*.[54] compared five different classification approaches to obtain the best performing method, whereas in another study only two methods were trialled.[58]

Image datasets used in the selected studies

Figure 4 illustrates the dataset sizes (for training and validation) and spatial dimensions of images (in pixels) for the selected studies.

There is apparent variability in sample sizes, with training samples ranging from 8 to 216 images and validation samples ranging from 0 to 208 images. The overall dataset size (including both training and test samples) ranges from 40 to 270 images (mean ~139 images).

The spatial dimensions of images ranged from 262,144 to 10,890,000 pixels; this excludes three studies where the image dimensions were not described[52,56,62]. Although the image sizes are measured in pixels, the actual size of the tissue sample (in microns) will differ due to the varying resolutions of different scanners and the magnification level chosen for the images. However, in five studies[54,55,57,58,62] the dataset samples were obtained from the same centre and in two studies[54,58] the same dataset was used.

Quality appraisal assessment

The quality of selected studies was assessed using a tailored QUADAS-2 tool. The overall score for each study was determined by combining the number of satisfied criteria, with a higher score representing higher quality evidence (Table 3). There was considerable variability in the methodological quality of included studies. Baik *et al.*[52] scored the highest for including the use of a separate validation and test set for optimal model selection and evaluation[52], whereas Sun *et al.*[60] scored the lowest across the 13 areas of assessment, largely due to the description of their approach without the use of reasonable dataset and results[60]. In all applicable studies, except for Lorsakul *et al.*[56], the methods and intermediate results were clearly presented. Rahman *et al.*[59] was the only study to have used a multi-centric dataset[59].

DISCUSSION

In order to safely and effectively implement automated AI-based methods in diagnostic and clinical practice, it is vital to validate these algorithms using a robust and fair experimental setup. This setup should include a clinically representative dataset and suitable evaluation metrics for validation.

The ideal dataset should represent clinical practice and take into account the whole tissue section. Tissue samples from multiple centres will enable greater diversity and biological variance through inclusion of cases from different geographical locations, patient populations and demographics. Furthermore, the ground truth should include meticulous annotations from multiple pathologists to minimise subjectivity and take into

account inter-pathologist variation. In this review, seven studies described more than one expert to be involved in providing ground truth, however it is not clear whether these refer to experienced pathologists, trainees, non-clinical researchers or other allied healthcare professionals. In most cases, the number of experts involved has also not been clearly stated. Furthermore, multi-centric data and WSI were used in only two studies (Rahman *et al*.[59] and Lorsakul *et al*[56] respectively). The majority of studies therefore used uni-centric datasets mostly comprising smaller sub-images within WSI, which may have introduced bias and would offer limited applicability.

In supervised ML methods, the ideal dataset should be divided into three groups for 1) model training 2) optimal model selection and 3) validation or evaluation. In this review, most of the assessed studies used the same dataset for both optimal model selection and evaluation. This indicates a high risk of bias which is likely to have contributed to the high accuracy rates (ranging from 79-100% across all studies). This bias could have been easily avoided by dividing these datasets into the three defined sub-sets or adding more unseen cases to the test and validation sets. The overall size of a dataset mainly depends on the type of AI method used. Traditional AI methods require small datasets whereas modern machine/deep learning methods require a larger dataset for model training. This concept is also known as the model generalisability and is concerned with the replication of model accuracy when applied to a new and diverse cohort of cases.

Most of the reviewed studies used AI methods which were regarded as the state-of-the-art at the time of publication and described these in sufficient detail to ensure reproducibility. Various evaluation metrics were utilised (accuracy, sensitivity, specificity etc.) to report the overall performance on the test set. However, in one study (Krishnan et al.[55]) the performance of individual images was measured, which makes it difficult to gauge the overall average performance of the entire test set. In four studies[53,54,58,61] a direct comparison of proposed techniques has been made to existing methods which should give some credibility to their proposed techniques. However, only three studies[53,54,61] compared their methods with the best performing methods at the time of publication.

Most of the selected studies were published before 2015, therefore, the methods employed were mainly classical ML methods[32,33,36]. This was somewhat surprising as the ML and AI fields have significantly progressed in the last decade, resulting in the development of numerous state-of-the-art algorithms for different real-world problems such as object detection in natural images[39,40,41], human voice recognition[42] and natural language processing[43]. These methods have multiple applications including medical image analysis which have been used to reliably predict diagnosis[18], mutational status[44] and treatment response[45] in a range of malignancies including breast, lung and colorectal cancers. However, our review shows that these latest AI methods have not been applied for detection of head and neck lesions, despite the ever-increasing global incidence and poor prognosis of HNC.

Our review highlights that a huge opportunity (and need) for medical image analysis and computational pathology researchers to develop novel methods to aid HNC diagnosis using modern AI approaches such as deep learning. Early work in this field appears to show potential for reliable detection between normal, potentially premalignant and cancerous lesions from histology WSI using classical classification methods[46,47]. Customised deep learning techniques have been used for segmentation of the epithelium[48,49] and cell segmentation has shown successful morphological analysis in HNC[50,51].

CONCLUSION

This review provides early evidence to support application of supervised ML methods as an aid to detection and grading in a limited number and types of OPMD. Furthermore, there is limited evidence exploring the use of AI to aid diagnosis of other potentially premalignant and cancerous head and neck lesions. Having said

this, most of the described AI/ML methods have the potential be modified for application to other clinical sites, including other head and neck lesions. The overall performance of the AI methods appears comparable to conventional light microscopic histopathological assessment but with added advantages of a faster, objective and reproducible evaluation. However, integration of these methods in the digital pathology workflow requires comprehensive evaluation of each method based on large multi-centric datasets. Future avenues include the use of deep learning methods for development of digital biomarkers and discovery of novel predictive features which will aid early detection of HNC and improve patient stratification. Ultimately, this will aid the development of targeted, patient-specific diagnostics and therapeutics to reduce HNC associated mortality.

Disclaimer: The content of this article represents the personal views of the authors and does not represent the views of the authors' employers and associated institutions. Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/ World Health Organization.

Acknowledgements

We thank Anthea Tucker (University of Sheffield, UK) for her assistance with the database searches.

Figure Captions:

Figure 1 – PRIMSA flowchart (diagram adapted from PRISMA group, 2009[28]) demonstrating the study selection process. (Colour not required)

Figure 2: Hierarchical classification of the methods used in the selected studies. (Colour not required)

Figure 3: A bar chart representing the frequency of different methods used in the selected studies. The colours correspond to the individual studies (as per key on right). (Colour required)

Figure 4: Graph demonstrating dataset sizes for selected studies. The area of the circle represents the spatial dimensions (in pixels) of the images used within the dataset whereas horizontal and vertical axis represent the number of images used for training and validation, respectively. Studies in which the image dimension is not provided have been marked as 'unknown'. (Colour required)

Table captions:

Table 1 – Glossary of relevant computational pathology terms. (Colour not required)

Table 2 - Summary of findings for selected studies. (Colour not required)

Table 3 – Quality assessment of the selected studies using modified QUADAS-2 tool. The '✔' demonstrates a favourable response to the question and the 'X' demonstrates an unfavourable response to the question. The overall score reflects the quality and risk of bias for each study. (Colour not required)

TERM	DESCRIPTION
Artificial Intelligence (AI)	A branch of computer science concerned with building smart machines that can perform tasks which typically require human intelligence.
Machine Learning (ML)	The ability for machines to 'learn' information and patterns directly from data without being programmed explicitly.
Supervised Learning	Training of ML algorithms from labelled (e.g. annotations) input and output data.
Unsupervised Learning	Training of ML algorithms by mining and extracting hidden patterns from input data that has not been labelled.
Deep Learning (DL)	DL is a subfield of ML in which algorithms learn from input data through example (without supervision).
Neural Network	A highly structured set of algorithms which models the brain's neural network system (deep learning) designed to recognise patterns from input data.
Whole slide image (WSI)	A high-resolution microscopy image of human tissue section.
F1 Score	A statistical analysis of binary classification to measure the accuracy of a test, considering the weighted average of the precision (p) and recall (r) of the test to compute the overall score.
Precision (p)	The number of correct positive results divided by the number of all positive results returned by the classifier
Recall (r)	The number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive
Classification	A ML technique which categorises a set of data (structured or unstructured) into classes based on certain attributes.
Patch extraction	Deconstruction of a WSI into smaller pixelated patches known as 'sub-images'.
Semantic segmentation	The process of dividing WSI's into regions of interest and clustering data into distinct groups based on similarities.

Table 1

Study	Study Reference	Study Aims	AI/ML Methods and Image Dataset	Reported Diagnostic Performance (%)	Risk of Bias
1	Baik *et al.*[52]	Classification of oral precancerous lesions into low and high-risk groups of progression into OSCC using a nuclear phenotypic score.	**Methods:** ○ Random Forest **Datasets:** ○ Training: 62 ○ Test: 71	○ Sensitivity: 78% ○ Specificity: 71%	9
2	Das *et al.*[53]	Segmentation of OSCC histology images into epithelial, sub-epithelial, and keratin layers. Detection of keratin pearls from the segmented keratin layer.	**Methods:** ○ Random Forest ○ Neural Network **Datasets:** ○ Training: 80 ○ Test: 20	○ Sensitivity 97.7% ○ Dice-coefficient 95.3%	4
3	Krishnan *et al.*[54]	Classification of oral precancerous lesions into normal, oral sub-mucous fibrosis without dysplasia, and oral sub-mucous fibrosis with dysplasia.	**Methods:** ○ Sugeno Fuzzy ○ Decision Tree ○ K-Nearest Neighbour ○ Gaussian Mixture Model ○ Neural Network **Datasets:** ○ Training: 158 ○ Test: 0	○ Sensitivity: 94.5% ○ Specificity 98.8%	7
4	Krishnan *et al.*[55]	Classification of oral precancerous lesions into normal and oral submucous fibrosis through segmentation of collagen fibres in the subepithelial connective tissue.	**Methods:** ○ Bayesian Classifier ○ Support Vector Machine ○ Neural Network **Datasets:** ○ Training: 89 ○ Test: 30	○ Accuracy: 91.70%	2
5	Lorsakul *et al.*[56]	Spatial quantification of brightfield-multiplex immunohistochemistry stained imaging for epithelial tumour cells and carcinoma-associated fibroblasts in tumour-associated stroma, through detection and classification of cells and segmentation of fibroblasts.	**Methods:** ○ Random Forest ○ L1-Logistic Regression ○ Support Vector Machine **Datasets:** ○ Training: 135	○ Accuracy: 91.64%	2

			○ Test: 35		
6	Krishnan *et al.*[57]	Segmentation of sub-epithelial connective tissue cells followed by cell classification into normal and oral submucous fibrosis tissue.	**Methods:** ○ Support Vector Machine **Datasets:** ○ Training: 20 ○ Test: 20	○ Sensitivity: 90.46% ○ Specificity: 87.54% ○ Accuracy: 88.89%	3
7	Krishnan *et al.*[58]	Segmentation of the epithelial layer in normal and oral sub-mucous fibrosis histological images.	**Methods:** ○ Watershed ○ Otsu **Datasets:** ○ Training: 158 ○ Test: 0	Standard performance measures not reported.	4
8	Rahman *et al.*[59]	Classification of oral histology images into normal and oral squamous cell carcinoma.	**Methods:** ○ Support Vector Machine **Datasets:** ○ Training: 216 ○ Test: 54	○ Accuracy: 100%	5
9	Sun *et al.*[60]	Segmentation of tumour in anti-CD34 antibody stained oral cancer histology images.	**Methods:** ○ Clustering **Datasets:** ○ Training: 8 ○ Test: 208	Standard performance measures not reported.	1
10	Fouad *et al.*[61]	Segmentation of oropharyngeal cancer tissue into epithelial and stromal regions.	**Methods:** ○ Clustering **Datasets:** ○ Training: 10 ○ Test: 45	F1-Score: 81%	5
11	Krishnan *et al.*[62]	Classification of oral premalignant lesions into normal, oral sub-mucous fibrosis with atrophy.	**Methods:** ○ Otsu ○ Linear Discriminant Analysis ○ Neural Network **Datasets:** ○ Training: 84 ○ Test: 28	○ Sensitivity: 92.31% ○ Specificity: 100	3

Table 2

Quality Assesment Questions	Study Number										
	1	2	3	4	5	6	7	8	9	10	11
Does the dataset include the complete digitised biopsy section or a complete resection?	✖	✖	✖	✖	✔	✖	✖	✖	✖	✖	✖
Does the dataset consist of more 100 samples?	✔	✖	✖	✖	✖	✖	✖	✔	✖	✖	✖
If not, does it consist of more than 50 unique samples?	✔	✖	✔	✖	✖	✖	✖	✔	✖	✔	✖
Is the test dataset separate to the training and validation datasets?	✔	✖	✖	✖	✖	✖	---	✖	✖	✖	✖
Is the biopsy case selection representative of the condition being assessed in the study?	✔	✖	✔	✖	✖	✖	✔	✖	✖	✖	✖
Does the study involve more than one pathologist for annotations?	✔	✖	✔	✖	✔	✔	✔	✔	✖	✖	✔
Is the dataset multi-centric?	✖	✖	✖	✖	✖	✖	✖	✔	✖	✖	✖
Does the study use an independent reference/test set?	✔	✖	✖	✖	✖	✖	---	✖	✖	✖	✖
Does the study use an independent validation set for optimal model selection?	✔	✖	✖	✖	✖	✖	---	✖	✖	✖	✖
Does the study fairly compare the outcomes of the AI methods to the existing methods?	✖	✔	✔	✖	✖	✖	✔	✖	✖	✔	✖
Are the compared methods state-of-the-art at the time of the publication of the article?	✖	✔	✔	✖	✖	✖	✖	✖	✖	✔	✖
Was the method described in sufficient detail to reproduce the presented results?	✔	✔	✔	✔	✖	✔	✔	✔	✔	✔	✔
Were the results of the intermediate stages reported?	✔	✔	✔	✔	---	✔	---	---	---	✔	✔
Overall Score	9	4	7	2	2	3	4	5	1	5	3

Table 3

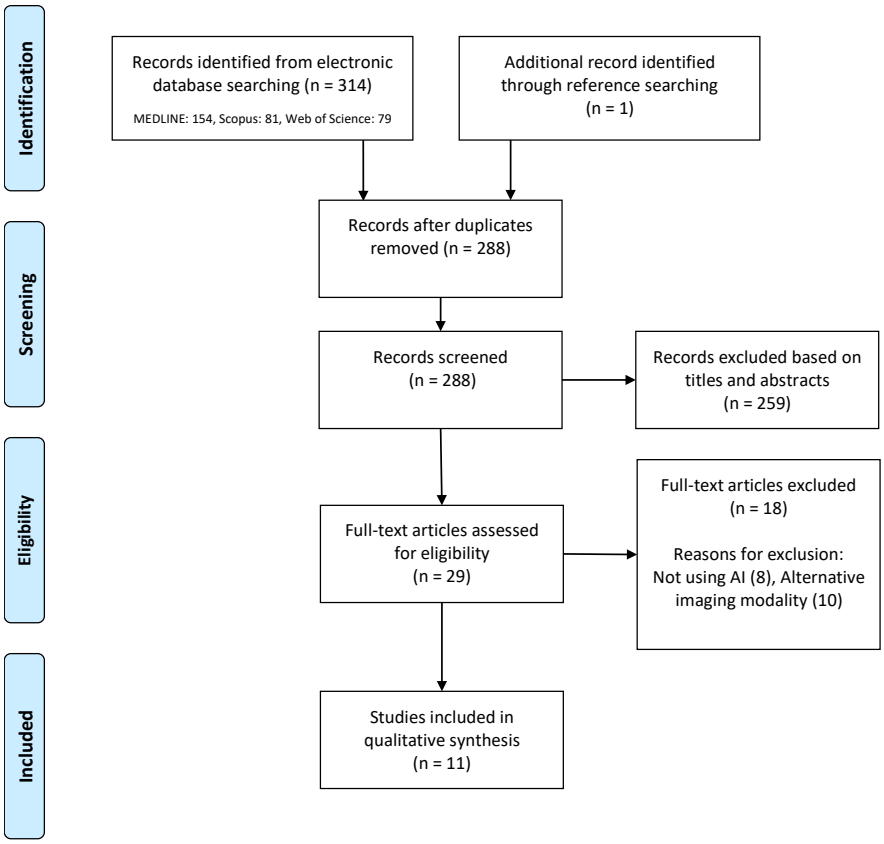

Figure 1

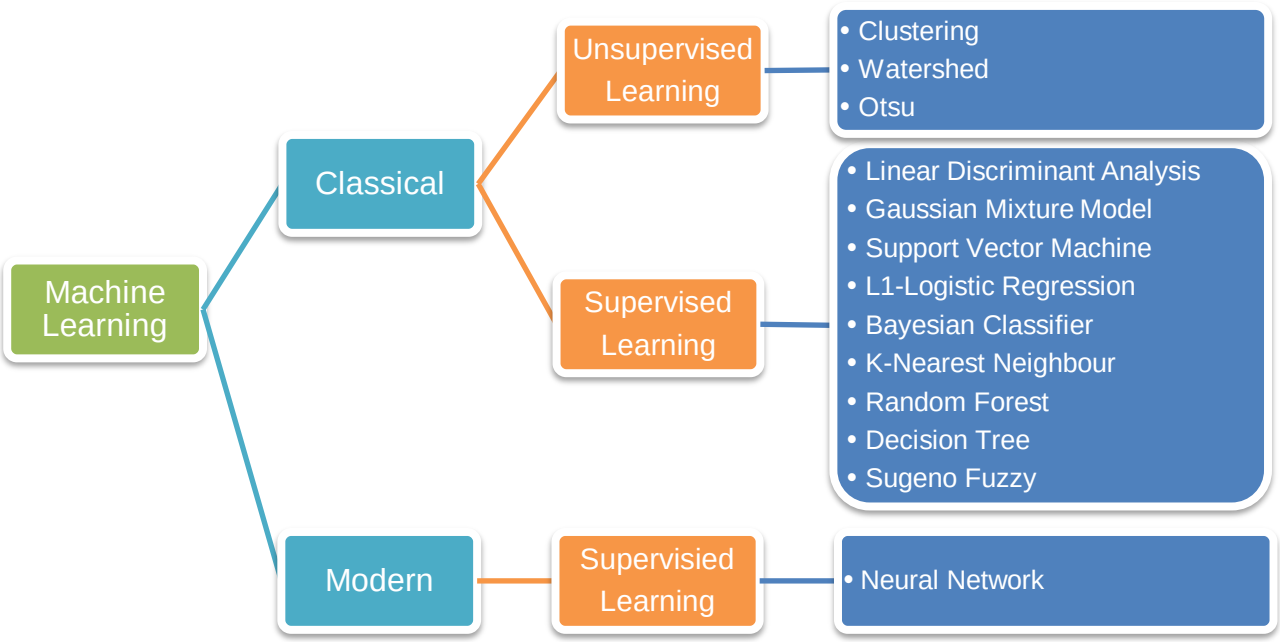

Figure 2

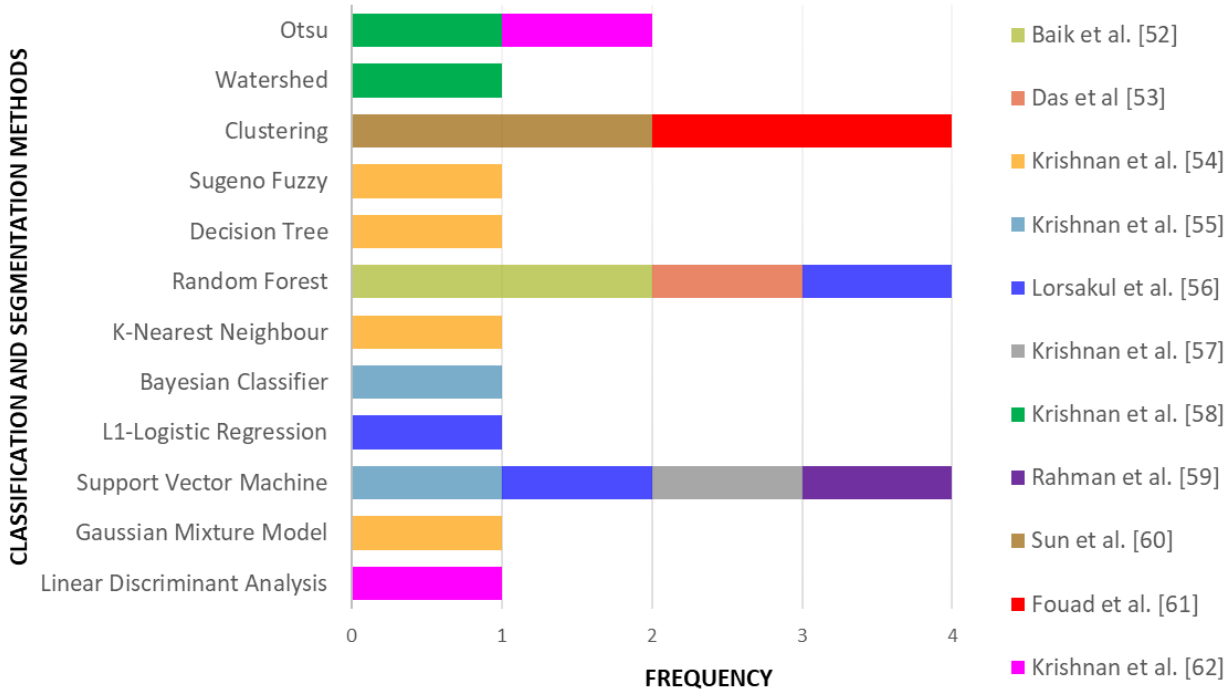

Figure 3

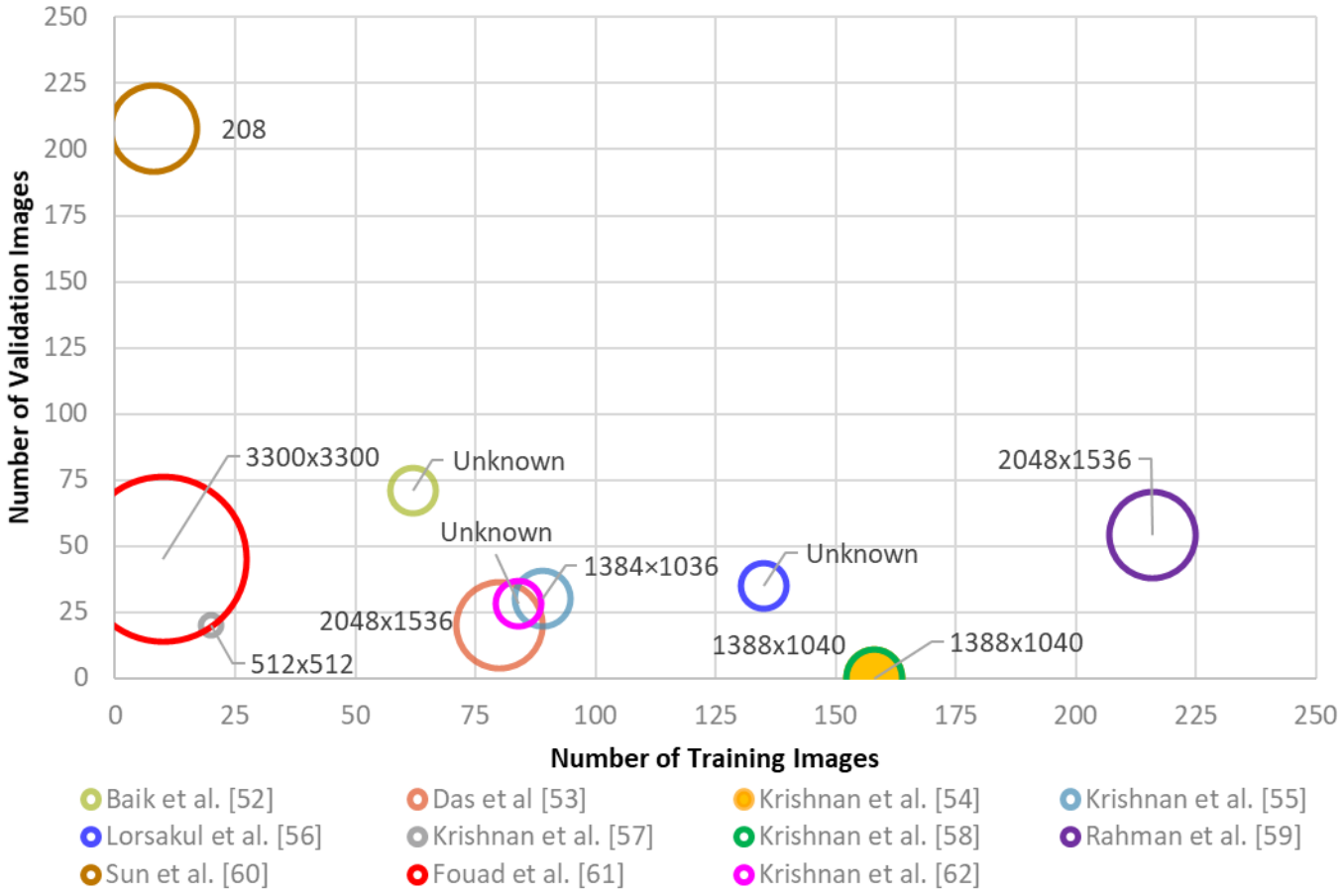
Number of Validation Images
Number of Training Images
208
3300x3300
Unknown
Unknown
1384×1036
2048x1536
512x512
Unknown
1388x1040
1388x1040
2048x1536
Baik et al. [52]
Das et al [53]
Krishnan et al. [54]
Krishnan et al. [55]
Lorsakul et al. [56]
Krishnan et al. [57]
Krishnan et al. [58]
Rahman et al. [59]
Sun et al. [60]
Fouad et al. [61]
Krishnan et al. [62]

Figure 4

References

[1] World Health Organization, IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, International Agency for Research on Cancer. Betel-quid and areca-nut chewing and some areca-nut-derived nitrosamines. IARC; 2004.

[2] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Alcohol consumption and ethyl carbamate. IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. 2010;96:3.

[3] International Agency for Research on Cancer. Solar and ultraviolet radiation. IARC monographs on the evaluation of carcinogenic risks to humans. 1992; 55..

[4] IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Human immunodeficiency viruses and human T-cell lymphotropic viruses. IARC monographs on the evaluation of carcinogenic risks to humans. 1996.

[5] Papillomaviruses H. IARC monographs on the evaluation of carcinogenic risks to humans. Lyon, France: IARC. 2011.

[6] Shotelersuk K, Khorprasert C, Sakdikul S, Pornthanakasem W, Voravud N, Mutirangura A. Epstein-Barr virus DNA in serum/plasma as a tumor marker for nasopharyngeal cancer. Clinical Cancer Research. 2000 Mar 1;6(3):1046-51.

[7] Shaw R, Beasley N. Aetiology and risk factors for head and neck cancer: United Kingdom National Multidisciplinary Guidelines. The Journal of Laryngology & Otology. 2016 May;130(S2):S9-12.

[8] Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA: a cancer journal for clinicians. 2011 Mar;61(2):69-90.

[9] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA Cancer J Clin 2020; 70:7.

[10] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2018; 68:394.

[11] Cancer Research UK. Cancer Incidence in the UK in 2011.

[12] Liao LJ, Hsu WL, Lo WC, Cheng PW, Shueng PW, Hsieh CH. Health-related quality of life and utility in head and neck cancer survivors. BMC cancer. 2019 Dec;19(1):425.

[13] Northern Ireland Cancer Registry, Queens University Belfast, Incidence by stage 2010-2014. Belfast: NICR; 2016.

[14] Warnakulasuriya S, Reibel J, Bouquot J, Dabelsteen E. Oral epithelial dysplasia classification systems: predictive value, utility, weaknesses and scope for improvement. Journal of Oral Pathology & Medicine. 2008 Mar;37(3):127-33.

[15] Lumerman H, Freedman P, Kerpel S. Oral epithelial dysplasia and the development of invasive squamous cell carcinoma. Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, and Endodontology. 1995 Mar 1;79(3):321-9.

[16] Kujan O, Khattab A, Oliver RJ, Roberts SA, Thakker N, Sloan P. Why oral histopathology suffers inter-observer variability on grading oral epithelial dysplasia: an attempt to understand the sources of variation. Oral oncology. 2007 Mar 1;43(3):224-31.

[17] Mehlum CS, Larsen SR, Kiss K, Groentved AM, Kjaergaard T, Möller S et al. Laryngeal precursor lesions: Interrater and intrarater reliability of histopathological assessment. The Laryngoscope. 2018 Oct;128(10):2375-9.

[18] Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama. 2017 Dec 12;318(22):2199-210.

[19] Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Computational and structural biotechnology journal. 2015 Dec 31;13:8-17

[20] Graham, S. & Rajpoot, N. SAMS-NET: Stain-aware multi-scale network for instance-based nuclei segmentation in histology images. In Biomedical Imaging (ISBI 2018), 2018 IEEE 15th International Symposium on, 590–594 (IEEE, 2018).

[21] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015 May;521(7553):436-44.

[22] Awan R, Sirinukunwattana K, Epstein D, Jefferyes S, Qidwai U, Aftab Z et al. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. Scientific reports. 2017 Dec 4;7(1):16852.

[23] Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature Reviews Clinical Oncology. 2019 Nov;16(11):703-15.

[24] Zormpas-Petridis K, Failmezger H, Raza SE, Roxanis I, Jamin Y, Yuan Y. Superpixel-based Conditional Random Fields (SuperCRF): Incorporating global and local context for enhanced deep learning in melanoma histopathology. Frontiers in oncology. 2019;9:1045.

[25] Wang S, Yang DM, Rong R, Zhan X, Fujimoto J, Liu H et al. Artificial Intelligence in Lung Cancer Pathology Image Analysis. Cancers. 2019 Nov;11(11):1673.

[26] Sirinukunwattana K, Shan e Ahmed Raza, Tsang YW, Snead DR, Cree IA, Rajpoot NM. Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE Trans. Med. Imaging. 2016 May 1;35(5):1196-206.

[27] Veta M, Van Diest PJ, Willems SM, Wang H, Madabhushi A, Cruz-Roa A et al.. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Medical image analysis. 2015 Feb 1;20(1):237-48.

[28] Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 2010;8:336–41.

[29] Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Annals of internal medicine. 2011 Oct 18;155(8):529-36.

[30] Yen J, Langari R, Zadeh LA, editors. Industrial applications of fuzzy logic and intelligent systems. IEEE; 1995 Apr 1.

[31] Quinlan JR. Induction of decision trees. Machine learning. 1986 Mar 1;1(1):81-106.

[32] Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.

[33] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992 Aug 1;46(3):175-85.

[34] Rish I. An empirical study of the naive Bayes classifier. InIJCAI 2001 workshop on empirical methods in artificial intelligence 2001 Aug 4 (Vol. 3, No. 22, pp. 41-46).

[35] Mika S, Ratsch G, Weston J, Scholkopf B, Mullers KR. Fisher discriminant analysis with kernels. InNeural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468) 1999 Aug 25 (pp. 41-48). Ieee.

[36] Suykens JA, Vandewalle J. Least squares support vector machine classifiers. Neural processing letters. 1999 Jun 1;9(3):293-300.

[37] Bilmes JA. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute. 1998 Apr 21;4(510):126.

[38] Fine TL. Feedforward neural network methodology. Springer Science & Business Media; 2006 Apr 6.

[39] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

[40] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2818-2826).

[41] Chollet F. Xception: Deep learning with depthwise separable convolutions. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 1251-1258).

[42] Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. InProceedings of the 23rd international conference on Machine learning 2006 Jun 25 (pp. 369-376).

[43] Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine. 2018 Jul 20;13(3):55-75.

[44] Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH et al.. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature medicine. 2019 Jul;25(7):1054-6.

[45] Shaban M, Khurram SA, Fraz MM, Alsubaie N, Masood I, Mushtaq S et al.. A novel Digital Score for Abundance of Tumour Infiltrating Lymphocytes predicts Disease free Survival in oral Squamous cell carcinoma. Scientific reports. 2019 Sep 16;9(1):1-3.

[46] Graham S, Shaban M, Qaiser T, Koohbanani NA, Khurram SA, Rajpoot N. Classification of lung cancer histology images using patch-level summary statistics. InMedical Imaging 2018: Digital Pathology 2018 Mar 6 (Vol. 10581, p. 1058119). International Society for Optics and Photonics.

[47] Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. InProceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 2424-2433).

[48] Fraz MM, Khurram SA, Graham S, Shaban M, Hassan M, Loya A et al. . FABnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer. Neural Computing and Applications. 2019 Nov 3:1-4.

[49] Chan L, Hosseini MS, Rowsell C, Plataniotis KN, Damaskinos S. Histosegnet. Semantic segmentation of histological tissue type in whole slide images. In Proceedings of the IEEE International Conference on Computer Vision 2019 (pp. 10662-10671).

[50] Raza SE, Cheung L, Shaban M, Graham S, Epstein D, Pelengaris S et al. . Micro-Net: A unified model for segmentation of various objects in microscopy images. Medical image analysis. 2019 Feb 1;52:160-73.

[51] Graham S, Vu QD, Raza SE, Azam A, Tsang YW, Kwak JT et al.. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis. 2019 Dec 1;58:101563.

[52] Baik J, Ye Q, Zhang L, Poh C, Rosin M, MacAulay C et al.. Automated classification of oral premalignant lesions using image cytometry and random forests-based algorithms. Cellular Oncology. 2014 Jun 1;37(3):193-202.

[53] Das DK, Bose S, Maiti AK, Mitra B, Mukherjee G, Dutta PK. Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis. Tissue and Cell. 2018 Aug 1;53:111-9.

[54] Krishnan MM, Venkatraghavan V, Acharya UR, Pal M, Paul RR, Min LC et al. Automated oral cancer identification using histopathological images: a hybrid feature extraction paradigm. Micron. 2012 Feb 1;43(2-3):352-64.

[55] Krishnan MM, Shah P, Chakraborty C, Ray AK. Statistical analysis of textural features for improved classification of oral histopathological images. Journal of medical systems. 2012 Apr 1;36(2):865-81.

[56] Lorsakul A, Andersson E, Harring SV, Sade H, Grimm O, Bredno J. Automated wholeslide analysis of multiplex-brightfield IHC images for cancer cells and carcinoma-associated fibroblasts. InMedical Imaging 2017: Digital Pathology 2017 Mar 1 (Vol. 10140, p. 1014007). International Society for Optics and Photonics.

[57] Krishnan MM, Pal M, Bomminayuni SK, Chakraborty C, Paul RR, Chatterjee J et al. Automated classification of cells in sub-epithelial connective tissue of oral sub-mucous fibrosis—An SVM based approach. Computers in biology and medicine. 2009 Dec 1;39(12):1096-104.

[58] Krishnan MM, Choudhary A, Chakraborty C, Ray AK, Paul RR. Texture based segmentation of epithelial layer from oral histological images. Micron. 2011 Aug 1;42(6):632-41.

[59] Rahman TY, Mahanta LB, Chakraborty C, Das AK, Sarma JD. Textural pattern classification for oral squamous cell carcinoma. Journal of microscopy. 2018 Jan;269(1):85-93.

[60] Sun YN, Wang YY, Chang SC, Wu LW, Tsai ST. Color-based tumor tissue segmentation for the automated estimation of oral cancer parameters. Microscopy Research and Technique. 2010 Jan;73(1):5-13.

[61] Fouad S, Randell D, Galton A, Mehanna H, Landini G. Unsupervised morphological segmentation of tissue compartments in histopathological images. PloS one. 2017;12(11).

[62] Mookiah MR, Shah P, Chakraborty C, Ray AK. Brownian motion curve-based textural classification and its application in cancer diagnosis. Analytical and quantitative cytology and histology. 2011 Jun;33(3):158-68.

Appendices

Appendix 1 - MEDLINE via OVID search strategy

1. artificial intelligence.mp. or exp Artificial Intelligence/
2. machine learning.mp. or Machine Learning/
3. deep learning.mp. or Deep Learning/
4. Image Processing, Computer-Assisted/ or automated detection.mp. or Diagnosis, Computer-Assisted/
5. "Neural Networks (Computer)"/ or neural networks.mp.
6. automated image analysis.mp.
7. digital image analysis.mp.
8. 1 or 2 or 3 or 4 or 5 or 6 or 7
9. Mouth Neoplasms/ or oral epithelial dysplasia.mp. or Leukoplakia, Oral/
10. oral leukoplakia.mp.
11. oral neoplasm.mp.
12. oral precancer.mp.
13. oral cancer.mp.
14. "head and neck cancer".mp. or "Head and Neck Neoplasms"/
15. "head and neck malignancy".mp.
16. 9 or 10 or 11 or 12 or 13 or 14 or 15
17. Diagnosis/ or diagnosis.mp.
18. diagnostic performance.mp.
19. 17 or 18
20. 8 and 16 and 19
21. limit 20 to (english language and humans and last 10 years)